A Machine Learning Approach for Predicting Antibody Properties

Oche A Egaji
The Centre of Excellence in Mobile and Emerging Technologies
University of South Wales, Pontypridd, United Kingdom
alexander.egaji@southwales.ac.uk

Seamus Ballard-smith
The Centre of Excellence in Mobile and Emerging Technologies
University of South Wales, Pontypridd, United Kingdom
seamus.ballard-smith@southwales.ac.uk

Ikram Asghar
The Centre of Excellence in Mobile and Emerging Technologies
University of South Wales, Pontypridd, United Kingdom
ikram.asghar@southwales.ac.uk

Mark Griffiths
The Centre of Excellence in Mobile and Emerging Technologies
University of South Wales, Pontypridd, United Kingdom
mark.griffiths@southwales.ac.uk

ABSTRACT

This paper used an amino acid location-based sequence encoding as a feature extraction techniques to identify single chains antibody molecules that bind to B-lymphocyte stimulator (BLyS) antigen. The data were manually derived from the European patent (EP2275449B1) text. The dataset was cleaned and made suitable for the machine learning models. The accuracy, precision and recall achieved across individual descriptors (Membrane and Soluble) for Logistic regression, KNN, KSVM, and Random Forest Tree was above 80%. However, it was much lower for the Naïve Bayes except for the precision score. The promising accuracy value achieved from such a minimal dataset has significant implications for the drug discovery process – this includes considerable savings in time and resources.

CCS Concepts

• Information systems → Information systems applications → Data mining → Data cleaning.

Keywords

Machine learning; Antigen, Antibody; Amino acid sequence, Infectious disease

1. INTRODUCTION

There is a global concern in the increase of drug resistance bacteria causing infectious diseases due there overuse. Infectious diseases are the leading cause of illness and death across the world. An infectious disease occurs when pathogens invade a host cell, starts to multiply and cause the abnormal functioning of the host cells. This is caused by the protein-protein interaction between the pathogen and the host cell[1] [2] [3]. The protein-protein interaction between the pathogen and host is the molecular basis of a disease, hence understanding this interaction will aid in the development of appropriate prevention, diagnosis and treatment approach [4].

Currently, these interactions are found by examining all possible protein-protein pairs with an element of guidance from previous experience. This approach is highly inefficient in both time and resources. Moreso, the reactions that can be detected via this approach only accounts for a small fraction of all possible reactions [2] [5]. For example, in the absence of antigen simulation, a human can make over 10^{12} potential antibodies molecules that can react to significant varying antigenic determinant [6]. Hence, advanced insight into understanding the protein-protein interaction between the pathogen and host cells with the aid of a predictive model would be a considerable advantage in improving the efficiency of the process. This can reduce the protein pairs to be experimentally tested, resulting in substantial savings in time and resources. The main barrier to creating such tools is in the acquisition of sufficient data to build up a dataset of a size that will allow successful training of the predictive model [7].

Hence, a key demonstrator in this paper was achieved using the amino acid sequence of antibody and antigen as the data input for the machine-learning model. This is a binary classification problem as the model predicts whether a single chains antibody molecule binds to B-lymphocyte stimulator (BLyS) antigen. The data were manually derived from the European patent (EP2275449B1) text. The data was augmented to boost its size and quality. The data was cleaned and made suitable for the machine-learning models.

The rest of this paper is organised as follows: section two and three contain the related work and a brief description of the machine-learning models, respectively. The research methodology and data overview are presented in section four and five, respectively. The sixth section contains the performance indicators and, the result and analysis. Finally, the conclusion is in the seventh section.

2. RELATED WORK

Recent research in this area [8], [9], [10], [11], [12] have focused on the use of deep learning models for predicting the pathogen-

ICICM 2020, August 12–14, 2020, Paris, France
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8770-5/20/08…$15.00
https://doi.org/10.1145/3418981.3418983

host interaction. Some of these researches include that carried out by Tian et al.; the authors used deep learning to predict compound-protein interaction over balanced and unbalanced datasets from the STITCH database. The author's claim that their approach outperformed the existing [9]. Wang et al. used deep learning to predict protein contacts using both the evolutionary coupling and sequence conservative information. The authors claim their model outperformed existing approach when tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398-membrane protein [10]. Wan et al. combined feature-embedding technique with deep learning to predict compound-protein interaction. The authors claim that the proposed model performed well when tested on the ChEMBL and BindingDB datasets [11].

Hamanaka et al. predict compound protein interaction using a deep learning model. The author's claim an accuracy of 98.2% [12]. Other authors such as [[13], [14]] have considered other machine learning approaches. Unterthiner et al. compared the performance of deep learning with other predictors, which includes two commercially used predictor. The authors claim that the deep learning model outperformed the other methods when tested on the ChEMBL dataset. Another technique using the comparative model was proposed by [1]. Kösesoya et al. used location-based encoding to predict pathogen-host interaction using several machine learning models. The authors claim their approach performed better than existing for decision tree (RF and J48) and Bayesian-based (BN and NB) classifiers when applied on the Bacillus Anthracis and Yersinia Pestis datasets as the pathogens and human protein as the host [13].

This paper used the amino acid sequence (Heavy and light chain) of antibody and antigen as the data input for the machine-learning model to identify single chains antibody molecules that bind to BLyS. The paper compared the performance of several machine-learning models such as Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest Tree, Naïve Bayesian and K-Nearest Neighbour (KNN).

3. MACHINE LEARNING ALGORITHM

Machine learning is a mathematical algorithm that enables a computer to learn from example data. The most common type of machine learning is supervised and unsupervised learning. The algorithm learns from an example labelled data when it is supervised learning and learns from unlabelled data when it is supervised learning.

The identification of single chains antibody molecules that bind to BLyS is a binary classification problem (supervised learning) [15]. The classification model will predict the protein-protein interaction between Blys and the host cells, i.e. the algorithm will attempt to predict a '1' when the BLys binds to a descriptor or a '0' otherwise. This paper will compare the performance classification models such as Support Vector Machine (SVM) [16], Logistic Regression [17], Random Forest Tree [18], Naïve Bayesian [19] and K-Nearest Neighbour (KNN) [20].

4. METHODOLOGY

The research methodology as shown in Figure 1, consists of the data collection/assembling phase, the data pre-processing, the train/test data split, feature scaling, data augmentation, dimension reduction, modelling and finally, the evaluation and analysis.

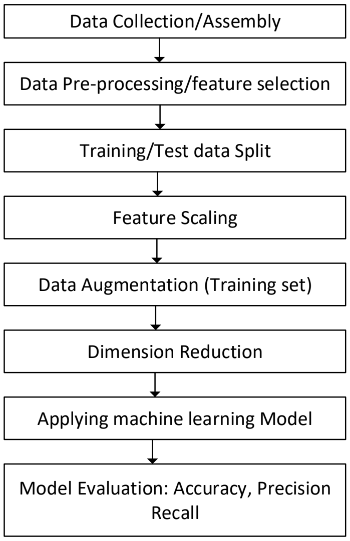

Data Collection/Assembly
Data Pre-processing/feature selection
Training/Test data Split
Feature Scaling
Data Augmentation (Training set)
Dimension Reduction
Applying machine learning Model
Model Evaluation: Accuracy, Precision Recall

Figure 1. Flow chart of research methodology

5. DATA OVERVIEW

The data were manually derived from the European patent (EP2275449B1) text. The authors of the patent formulated the data using phage display to identify single chains antibody molecules that bind to BLyS. These include single chains antibody molecules that bind to the soluble form of BLyS, single chains antibody molecules that bind the Membrane-bound form of BLyS, and single chains antibody molecules that bind to both the Soluble form and the Membrane-bound form of BLyS [21].

5.1 Data Preprocessing

The data consists of several protein sequences for a particular antibody. The outcome of the interaction between the protein sequence against the BLyS antigen for the four bond descriptors (Soluble and Membrane) is 1 when it binds or 0 otherwise. The amino acid sequence for Domain 1 (corresponds to the Heavy Chain) while that for Domain 2 (Corresponds to the Light chain). The Amino acids were encoded as integer values 1 through 24 as shown in

Overall, the dataset consists of 2103 unique antibodies with 298 features. The ratio of 'null' to 'positive' data for Soluble and Membrane are 246:1857 and 318:1785, respectively. The 'positive' data implies the sequence binds to the descriptor while the null implies otherwise.

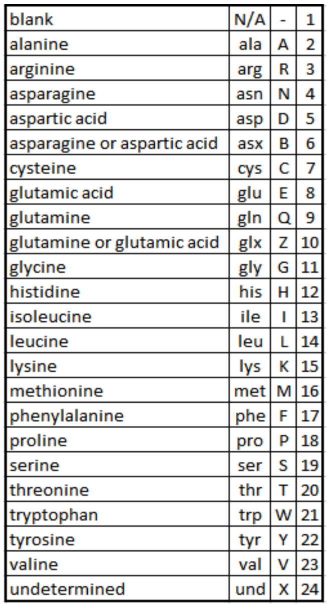
Further investigation of the datasets reveals that the values for 41 features were constant for all 2103 entries (40 of the unchanging features were blank - encoded as '1' and one feature tryptophan - encoded as '21'). All 41 features were removed from the datasets and the total feature left was 257. This is a small and unbalances dataset with the potential of causing a high variance in our predictive model. However, the availability of small dataset is a common phenomenon in drug discovery. To mitigate this problem, the 'null' dataset was upsampled using Synthetic Minority Over-sampling Technique (SMOTE) to provide an equal number of positive and negative data [22]. The data was split into 60% training and the remaining 40% for testing.

Table 1. Where 1 represents a blank in the sequence, 2 through 23 represents the 22 amino acids, and 24 represented any undetermined or X values.

Overall, the dataset consists of 2103 unique antibodies with 298 features. The ratio of 'null' to 'positive' data for Soluble and Membrane are 246:1857 and 318:1785, respectively. The 'positive' data implies the sequence binds to the descriptor while the null implies otherwise.

Further investigation of the datasets reveals that the values for 41 features were constant for all 2103 entries (40 of the unchanging features were blank - encoded as '1' and one feature tryptophan - encoded as '21'). All 41 features were removed from the datasets and the total feature left was 257. This is a small and unbalances dataset with the potential of causing a high variance in our predictive model. However, the availability of small dataset is a common phenomenon in drug discovery. To mitigate this problem, the 'null' dataset was upsampled using Synthetic Minority Over-sampling Technique (SMOTE) to provide an equal number of positive and negative data [22]. The data was split into 60% training and the remaining 40% for testing.

Table 1. Standard Amino acid data encoding



blank	N/A	-	1
alanine	ala	A	2
arginine	arg	R	3
asparagine	asn	N	4
aspartic acid	asp	D	5
asparagine or aspartic acid	asx	B	6
cysteine	cys	C	7
glutamic acid	glu	E	8
glutamine	gln	Q	9
glutamine or glutamic acid	glx	Z	10
glycine	gly	G	11
histidine	his	H	12
isoleucine	ile	I	13
leucine	leu	L	14
lysine	lys	K	15
methionine	met	M	16
phenylalanine	phe	F	17
proline	pro	P	18
serine	ser	S	19
threonine	thr	T	20
tryptophan	trp	W	21
tyrosine	tyr	Y	22
valine	val	V	23
undetermined	und	X	24

5.2 Dimension Reduction Algorithm

Factor analysis was performed on the remaining 257 features of the dataset to extract common variance and put them into a common factor score. Factor analysis is used to reduce a large number of variables into less number of factors. Applying Principal component analysis (PCA) [23] and using the Kaiser's criterion (eigenvalue > 1 rule) [24], the cumulative percentage of variance extracted to determine the number of factors. Out of the 256 potential factors, 6 have an eigenvalue > 1 with a cumulative variance of 99.709%. The eigenvalues are shown in Table **2**. After rotation the first factor accounts for (41.738%) of the variance, the second, third, fourth, fifth, sixth and seventh factor accounted for 24.766%, 18.866%, 10.596%, 3.169%, 0.573% and 0.29% of the variance respectively. The total variance described by the 1st to 6th factor is 99.709%, which is well above the acceptable limit of 40% [25]. Hence, PCA was applied to reduce the input 257 features to 6.

Table 2. Variance and Cumulative Percentage

Component Number	Extraction Sums of Squared Loadings		
	Actual eigenvalue from PCA	% of Variance	Cumulative %
1	107.267	41.738	41.738
2	63.649	24.766	66.504
3	48.485	18.866	85.370
4	27.233	10.596	95.966
5	8.145	3.169	99.136
6	1.473	0.573	99.709
7	0.749	0.291	100.000

6. RESULTS AND ANALYSIS

This result compares the performance of 6 classification models, namely: Logistic Regress, KNN, SVM, Naïve Bayesian and Random Forest Tree for the two predictors (Membrane and Soluble). The performance metric for the predictive model is accuracy, precision, recall and f-score. The accuracy consists of the ratio of correctly predicted observation to the total observation, and it is the commonly used predictive model evaluation metric. It is common practice to confuse high accuracy with better model performance; however, this can only be the case when the false positive and false negative are almost the same for a balanced dataset. A sample confusion matrix, as shown in Table 3, helps to illustrate the terms (true-positive, true-negatives, false-positives and false-negatives). The true-positives and true-negatives predictions are the correctly classified positive and negatives variables, respectively. The false-positives are the misclassified positive variables, and the false negatives are the misclassified negative variables.

Table 3. Confusion matrix

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Precision is the ratio of the correctly predicted positive observation of the total predicted positive observation. The precision is a good measure when the cost of false positive is high. While, the recall is a good measure when there is a high cost associated with false-negative, for example, in fraud detection or healthcare. The f-score is the weighted average of precision and recall. The precision, recall and f-score is a much better performance measure for a predictive model than accuracy, especially for an imbalanced dataset. The f-score reaches its best value at 1, and it worst at zero [26] [27].

The accuracy, Precision, Recall and f-score for Membrane and Soluble are shown in Table 4 and 5, respectively.

Table 4. Performance Metric: Membrane

Model	Accuracy	Precision	Recall	F-score
Logistic Regression	0.80	0.91	0.85	0.88
KNN	0.76	0.97	0.74	0.84
KSVM	0.81	0.91	0.86	0.88
Naive Bayes	0.44	0.94	0.37	0.53
Random Forest	0.83	0.91	0.89	0.89

Table 5. Performance Metric: Soluble

Model	Accuracy	Precision	Recall	F-score
Logistic Regression	0.96	0.98	0.97	0.98
KNN	0.93	0.99	0.92	0.96
SVM	0.95	0.98	0.96	0.97
Naive Bayes	0.48	0.98	0.42	0.59
Random Forest	0.97	0.99	0.97	0.98

The Naive Bayes model showed the worst performance, according to Table 4 and 5. It has low accuracy, recall, f-score but a high precision score. The Random forest tree is the best performing predictor based on the accuracy, precision, recall and f-score as compared to the other models. The performance metrics for the Random Forest Tree for Membrane and Soluble as shown in Table 4 and 5 are (Accuracy = 0.83, Precision = 0.91, recall = 0.89 and F1 score = 0.89) and (Accuracy = 0.97, Precision = 0.99, recall = 0.97and F1 score = 0.98), respectively. The precision scores for the Membrane and Soluble, implies that, out of all the positively predicted interaction, only 91 and 99%, respectively, truly bind to BLys. While, a recall of 0.89 and 0.97 for Membrane and Soluble means that out of all the antibodies that truly bind with BLys, the Random Forest tree correctly predicted 89 and 97%, respectively.

Overall, all the machine learning models performed bettered for soluble than membrane.

7. CONCLUSION

This paper uses an amino acid location-based sequence encoding as a feature extraction techniques to identify single chains antibody molecules that bind to BLyS. The data were manually derived from the European patent (EP2275449B1) text. Data augmentation was performed to increase the datasets for machine learning. The performance of machine learning models such as Logistic Regression, KNN, SVM, Naïve Bayesian and Random Forest Tree on the BLys dataset was discussed. PCA was applied to reduce the dimension of the features from 257 to 6. The performance metric for evaluating the models is the accuracy, precision, recall and f-score. The Random Forest Tree showed the best performance as compared to Logistic regression, KNN, KSVM and Naïve Bayesian for Membrane and Soluble, respectively.

The main barrier to creating/improving such tools is in the acquisition of sufficient data to build up a dataset of a size that will allow successful training of the predictive model. However, the high level of accuracy achieved from such a minimal dataset has significant implications for the drug discovery process in terms of considerable savings in time and resources.

8. ACKNOWLEDGMENTS

The authors will like to acknowledge the European Regional Development Fund (ERDF) and the Welsh Government for funding this study. Our gratitude goes to all members of the Centre of Excellence in Mobile and Emerging Technologies (CEMET), the University of South Wales for their contribution in various capacity in this study.

9. REFERENCES

[1] F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow, and A. Sali, "Host–pathogen protein interactions predicted by comparative modeling," *Protein Sci. Publ. Protein Soc.*, vol. 16, no. 12, pp. 2585–2596, Dec. 2007, doi: 10.1110/ps.073228407.

[2] K. Lage, "Protein-protein interactions and genetic diseases: The interactome," *Biochim. Biophys. Acta*, vol. 1842, no. 10, pp. 1971–1980, Oct. 2014, doi: 10.1016/j.bbadis.2014.05.028.

[3] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein Interactions and Disease," *PLoS Comput. Biol.*, vol. 8, no. 12, Dec. 2012, doi: 10.1371/journal.pcbi.1002819.

[4] S. Durmuş, T. Çakır, A. Özgür, and R. Guthke, "A review on computational systems biology of pathogen-host interactions," *Front. Microbiol.*, vol. 6, p. 235, 2015, doi: 10.3389/fmicb.2015.00235.

[5] L. Cai, Z. Pei, S. Qin, and X. Zhao, "Prediction of Protein-Protein Interactions in Saccharomyces cerevisiae Based on Protein Secondary Structure," in *2012 International Conference on Biomedical Engineering and Biotechnology*, May 2012, pp. 413–416, doi: 10.1109/iCBEB.2012.302.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, "The Generation of Antibody Diversity," *Mol. Biol. Cell 4th Ed.*, 2002, Accessed: Jul. 08, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK26860/.

[7] E. Nourani, F. Khunjush, and S. Durmuş, "Computational approaches for prediction of pathogen-host protein-protein interactions," *Front. Microbiol.*, vol. 6, p. 94, 2015, doi: 10.3389/fmicb.2015.00094.

[8] M. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Syst.*, vol. 8, no. 4, pp. 292–301, 2019.

[9] K. Tian, M. Shao, Y. Wang, J. Guan, and S. Zhou, "Boosting compound-protein interaction prediction by deep learning," *Methods*, vol. 110, pp. 64–72, 2016.

[10] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLoS Comput. Biol.*, vol. 13, no. 1, p. e1005324, 2017.

[11] F. Wan and J. Zeng, "Deep learning with feature embedding for compound-protein interaction prediction," *bioRxiv*, p. 086033, 2016.

[12] M. Hamanaka *et al.*, "CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning," *Mol. Inform.*, vol. 36, no. 1–2, p. 1600045, 2017.

[13] İ. Kösesoy, M. Gök, and C. Öz, "A new sequence based encoding for prediction of host–pathogen protein interactions," *Comput. Biol. Chem.*, vol. 78, pp. 170–177, 2019.

[14] T. Unterthiner *et al.*, "Deep learning for drug target prediction," *Work Represent Learn Methods Complex Outputs*, 2014.

[15] M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, “A Review of Classification Problems and Algorithms in Renewable Energy Applications,” *Energies*, vol. 9, no. 8, p. 607, Aug. 2016, doi: 10.3390/en9080607.

[16] T. Fletcher, “Support vector machines explained,” *Tutor. Pap.*, 2009.

[17] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, “An Introduction to Logistic Regression Analysis and Reporting,” *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: 10.1080/00220670209598786.

[18] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 157–175.

[19] X. Chai, L. Deng, Q. Yang, and C. X. Ling, “Test-Cost Sensitive Naive Bayes Classification,” in *Proceedings of the Fourth IEEE International Conference on Data Mining*, USA, Nov. 2004, pp. 51–58, Accessed: Jan. 14, 2020. [Online].

[20] H. binti Jaafar, N. binti Mukahar, and D. A. binti Ramli, “A methodology of nearest neighbor: Design and comparison of biometric image database,” in *2016 IEEE Student Conference on Research and Development (SCOReD)*, Dec. 2016, pp. 1–6, doi: 10.1109/SCORED.2016.7810073.

[21] “Espacenet - Home page.” https://worldwide.espacenet.com/ (accessed Jul. 08, 2019).

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[23] S. Maitra and J. Yan, “Principle component analysis and partial least squares: Two dimension reduction techniques for regression,” *Appl. Multivar. Stat. Models*, vol. 79, pp. 79–90, 2008.

[24] H. F. Kaiser, “The application of electronic computers to factor analysis,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 141–151, 1960.

[25] G. H. Danteman, *Principal Components Analysis, Quantitative Applications in the Social Sciences*. Sage publications, inc, 1989.

[26] K. Divya, P. M, and P. Pabitha, “Analysing the Competency of Various Decision Trees towards Community Formation in Multiple Social Networks,” in *2019 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2019, pp. 0099–0103, doi: 10.1109/ICCSP.2019.8698110.

[27] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” in *Australasian joint conference on artificial intelligence*, 2006, pp. 1015–1021.